Lab #8
Cluster Analysis

In this lab, you will be working with an R data set that was run on the Affymetrix human HGU95Av2 array. The microarray data are from genomic primary fibroblast cell lines and were generated for 46 samples: 23 human (*Homo sapien*), 11 bonobo (*Pan paniscus*), and 12 gorilla (*Gorilla gorilla*) donors. This is a publicly available dataset within the 'fibroEset' package in R. It should be noted that two identical human donor arrays are in this dataset. This data set is good for clustering and classification problems since there is a large difference in transcript profiles between all 3 species.

The analysis that you will conduct is based on clustering methods. The first problems require hierarchical clustering, while the last problems use spectral *k*-means clustering. We denote this as 'spectral' because instead of using the genes/probes as input into the clustering algorithm like the hierarchical clustering method, some form of spectral decomposition (e.g. PCA) is first computed and these eigenfunctions are used in the clustering algorithm. This method can be more useful than using the genes/probes in some cases where the variability is best summarized in a few components (or eigenfunctions).

1.) Load the fibroEset library and data set. Obtain the classifications for the samples.

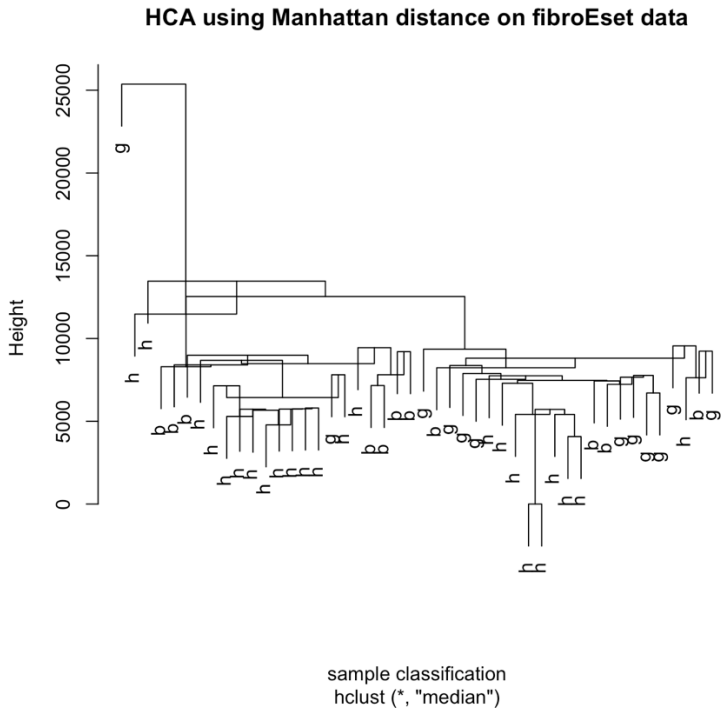
```
> library(fibroEset)
> data(fibroEset)
> dat <- fibroEset
> dat$species
 [1] b b b b b b b b b b b g g g g g g g g g g g g h h h h h
[29] h h h h h h h h h h h h h h h h h h
> dat <- as.data.frame(dat)
```

2.) Select a random set of 50 genes from the data frame, and subset the data frame.

```
> genes <- sample(dat, 50)
> ann <- dat$species
```

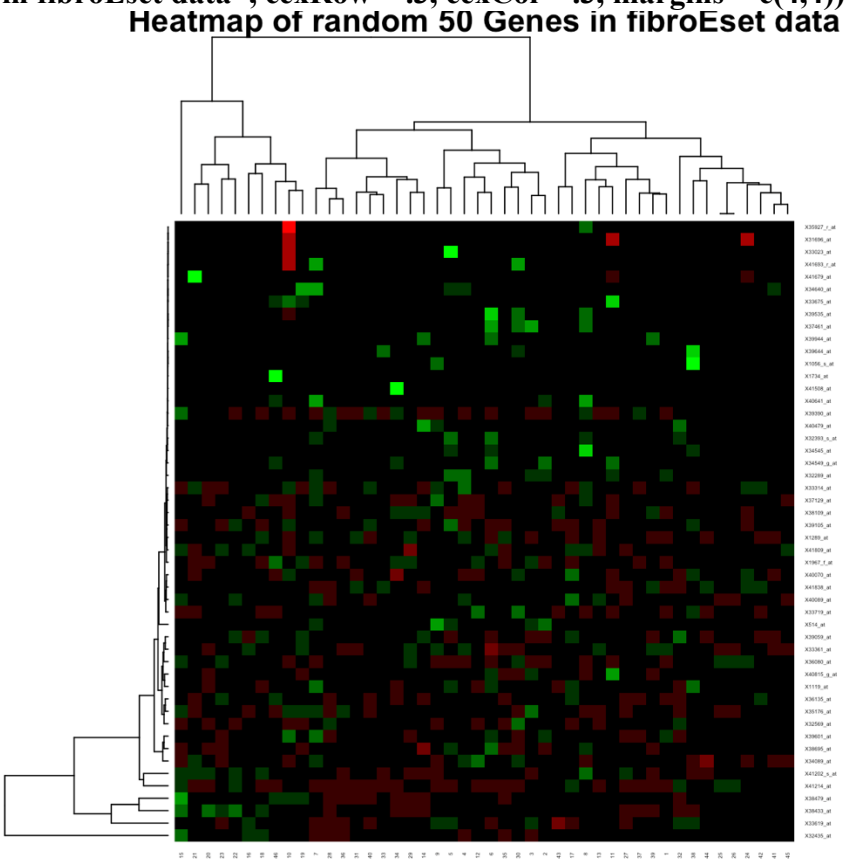
3.) Run and plot hierarchical clustering of the samples using manhattan distance metric and median linkage method. Make sure that the sample classification labels are along the x-axis. Title the plot.

> distance.matrix <- dist(genes, method="manhattan")
> gene.clust <- hclust(distance.matrix, method="median")
> plot(gene.clust, labels = ann, main = "HCA using Manhattan distance on fibroEset data", xlab = "sample classification")



4.) Now both run hierarchical clustering and plot the results in two dimensions (on samples and genes). Plot a heatmap with the genes on the y-axis and samples on the x-axis. Once again, make sure that the sample and genes labels are present. Title the plot.

> hm.rg <-
c("#FF0000","#CC0000","#990000","#660000","#330000","#000000","#00
0000","#0A3300","#146600","#1F9900","#29CC00","#33FF00")
> heatmap(as.matrix(t(genes)),col=hm.rg, main = "Heatmap of first 50 Genes in fibroEset data", cexRow = .3, cexCol = .3, margins = c(4,4))



5.) Calculate PCA on the samples and retain the first two components vectors (eigenfunctions). Calculate *k*-means clustering on these first two components with *k*=3.

> genes.pca <- prcomp(genes, cor=F)
> top2 <- genes.pca$x[,1:2]
> genes.pca.kmeans <- kmeans(top2, centers = 3, iter.max = 20)

6.) Plot a two-dimensional scatter plot of the sample classification labels, embedded with the first two eigenfunctions (from PCA). Color the labels with the color that corresponds to the predicted cluster membership. Make sure to label the axes and title the plot.

> ann <- as.character(ann) #turn ann into a characer vector to plot
> plot(top2, col = genes.pca.kmeans$cluster,cex=1, main = "K-means clustering of first 50 genes in fibroEset Data", xlab = "Data[,1]", ylab = "Data[,2]", pch=ann)